USING WORD TEMPORAL STRUCTURE IN HMM SPEECH RECOGNITION

L. Fissore ◇ and P. Laface ⋆ and F. Ravera ◇

◇ CSELT - Centro Studi e Laboratori Telecomunicazioni
Via G. Reiss Romoli 274 - I-10148 Torino, Italy
E-Mail fissore/ravera@cselt.stet.it
⋆ Dipartimento di Automatica e Informatica - Politecnico di Torino
Corso Duca degli Abruzzi 24 - I-10129 Torino, Italy
E-Mail laface@polito.it

ABSTRACT

Isolated word speech recognizers with fixed vocabularies are often used to provide vocal services through the telephone line. The paper illustrates a simple post-processing approach that allows the hypotheses produced by a Hidden Markov Model recognizer to be rescored taking into account the global temporal structure of the pronounced words. Our approach does not directly rely on state/word duration modeling. It models, instead, the global time variations of the spectral features of each word and their correlation in time: two important perceptual cues that are only partially exploited by standard HMMs.
This method has been evaluated using three isolated word speaker independent systems with vocabulary of different size and complexity. We show that, with minimal overhead, the recognition performance improves not only for small vocabulary recognition systems such as the isolated digit one, or for the recognition of 26 Italian spelling names, but also for a system with a 475 city name vocabulary included in a vocal service that provides information about the main railway connections.

1. INTRODUCTION

It is well known that one of the main deficiency of the classical HMMs is related to inadequate modeling of the duration of the acoustic events associated with each state. Since the probability of recursion to the same state is constant, the duration probability of the acoustic event associated with the state has an exponentially decreasing probability. This probability distribution of the durations does not correctly model the speech temporal structure.
To face this problem several techniques have been proposed that are recalled in [5]. These techniques rely on state duration modeling by means of discrete or continuous distributions that are more adequate to fit the temporal structure of speech, or they use the state duration as an additional information for rescoring the hypotheses produced by Viterbi decoding in a post-processing approach. The algorithms of the first type apply the state duration information within the decoding procedure, at the expense of a substantial increase of the computational complexity and memory occupation. The computational complexity of the algorithms of the second type is instead minimal. Since the duration information is not applied during the decoding procedure, the beam search strategy could sometimes eliminate correct word hypotheses. In most applications, however, the correct word is normally found within the top ten hypotheses, and a rescoring strategy based on the results obtained by means of different models, can be used to recover some errors.

A second weakness imputed to HMMs is the assumption that within each state the observation vectors are not correlated, while in reality the opposite is true. The information has be exploited to design more robust recognizer since it has been experimentally shown that spectral variations are discriminant features for similar sounds [2]. To encode within a state the dynamic features of the observation vectors several approaches have been proposed that can be summarized into four categories illustrated or referred in [3]:

- Addition to the observation vector of the first and second derivative of each component.
- Use of conditional probabilities within each state.
- Explicit modeling of the spectral vector correlation by means of linear prediction models associated to each state.

- Use of two-dimensional cepstral features.

The first solution is the most popular and effective, while for the second and third solutions one must deal with problems related to the parameter estimation accuracy and accept an increase of the decoding complexity.
As an alternative to the first one, the fourth solution is claimed to produce more robust spectral features with respect to cepstral coefficients including higher order derivatives [4].

All these solutions, however, do not account for global spectral variations. Thus, they are not able to avoid recognition errors deriving by an incorrect time warping. Many errors often occur, indeed, because a sequence of observations is decoded by a few states - typically adsorbing low energy frames - with high probability and duration. The other states, instead, are rapidly traversed because their distributions do not fit well the remaining observations. These errors, therefore, do not depend on the intrinsic confusion of acoustically similar words, rather, the lack of good duration modeling and an incorrect time warping produces word hypotheses that are loosely related to the acoustics of the correct word.

To account for the dynamic structure of the observation vectors, including both local and global variations, our approach does not directly rely on state/word duration modeling, rather it models the global time variations of the spectral features of each word and their correlation in time: two important perceptual cues that are only partially exploited by standard HMMs.
In particular, in this work we propose to rescore the probability produced by a conventional HMM system by means of the probability of a second very simple recognizer using word "temporal" models. The HMM system, using the first order derivative of the cepstral parameters and RASTA processing, takes care of the local variations, while in the second system, the global time spectral variations of a word are modeled by means of two-dimensional cepstral features [1].

2. CEPSTRAL-TIME MATRIX

A two-dimensional cepstral-time matrix [3] is the product of a Discrete Cosine Transform (DCT) performed on a sequence of T Mel-Frequency Cepstral Coefficient observation vectors along the time axis.

$$c_k(n,T) = \frac{1}{T}\sum_{t=1}^{T} mfcc_k(t) \cdot \cos\frac{(2t-1)\cdot n\cdot \pi}{2T} \quad (1)$$

where k, $1 \leq k \leq K$, refers to the $k-th$ observation vector component, t indicates the time frame, and n, $1 \leq n \leq N$, is the order of the DCT along the time axis. We are interested in the lower components (k and n) of the two-dimensional cepstral matrix of equation (1) because they encode the long-time variations of the spectral envelope [1, 4].

3. TEMPORAL MODELS TRAINING

To estimate the parameters of the temporal models of a word w, every utterance in the training set is processed to produce its two-dimensional cepstral matrix $c_k(n,T)$, where T is the time frame length of the utterance. Then, computing the mean ($\overline{\mu}_w(n)$) and variance ($\overline{\sigma}_w(n)$) vectors for each order n of the DCT, we obtain the word temporal model that consists of N K-dimensional Gaussian densities, one for each order of the DCT along the time axis:

$$\lambda(w) = \{\mathcal{N}(\overline{x}(n), \overline{\mu}_w(n), \overline{\sigma}_w(n))\} \qquad 1 \leq n \leq N$$

where $\overline{x}(n)$ is a K-dimensional DCT vector of order $n-th$.
During recognition, to score of an utterance according to a "temporal" model, we compute the log probability of the observation vectors DCT along the time axis, given the word model $\lambda(w)$, that is given by:

$$log P(DCT(O_1^T)|\lambda(w)) = \sum_{n=1}^{N} log\mathcal{N}(\overline{c}(n,T), \overline{\mu}_w(n), \overline{\sigma}_w(n)) \quad (2)$$

where $\overline{c}(n,T)$ is the $n-th$ column of the two-dimensional cepstral-time matrix $c_k(n,T)$.

It is worth noting that the overhead for the DCT computation is minimal, that a model includes only a single Gaussian density for each order of the DCT, and that no Dynamic Programming is performed to compute the log probability of Eq. 2. Thus, compared with the Viterbi decoding of continuous density mixture models, the memory occupation for the DCT and the complexity of the likelihood computation is negligible.

4. RESCORING

To decode a given utterance, first its endpoints are detected, and a HMM recognizer is activated that produces a set of word hypotheses with their associated log probabilities. Then, the two-dimensional cepstral matrix $c_k(n,T)$ is computed, and the temporal model

Table 1: Results comparing the baseline system and the rescoring approach

Vocabulary	Digits	Spelling	City names (manual EPD)	City names (automatic EPD)
Vocabulary Size	10	26	475	475
Test set size	5178	14079	14440	14440
HMM errors	65 (1.3%)	133 (1.0%)	682 (4.7%)	727 (5.1%)
Rescoring errors	53 (1.0%)	95 (0.7%)	459 (3.2%)	625 (4.4%)
Improvement %	18.4%	28.6%	32.7%	14.0%

score for these candidates is obtained by means of Eq. 2. Finally, the HMM word hypotheses are rescored according to the following steps:

1. Normalization of the HMM word candidate log probabilities.
   The HMM normalized score is obtained by

   $$hmm\_score(w) = (logP(O_1^T|\lambda(w)) - \mu)/\sigma$$

   where μ is the mean value and σ is the variance of the log probability $(logP(O_1^T|\lambda(w)))$ of all word candidates that have not been pruned out by the beam search. The maximum number of word candidates is limited to 10.

2. The same normalization is performed for the temporal model log probabilities to produce the temporal normalized score $t\_score(w)$.

3. The final score for each hypothesis is then computed by linear interpolation of the two scores:

   $$score(w) = \alpha \cdot hmm\_score(w) + (1.0 - \alpha) \cdot t\_score(w)$$

   and α is estimated using an independent evaluation set.

5. RESULTS

To test this approach we trained whole word HMMs for two small vocabulary tasks - digits and 26 Italian spelling names recognition - and 333 application dependent subword models for recognition of 475 city names. This latter vocabulary is used by a vocal service that provides information about the main railway connections. It is worth noting that the number of utterances for each word included in the database collected for training the city names allow the corresponding temporal model to be accurately estimated.

Table 1 summarizes, in its first three columns, the results obtained in a set of recognition experiments with the above mentioned vocabularies. For these experiments, 12 cepstral, 12 delta-cepstral coefficients and RASTA filtering for the HMMs feature vectors have been used, while $N = 8$ x $K = 8$ was the rank of the two-dimensional cepstral matrix for the temporal models and the interpolation factor α was set to 0.4 for the first two vocabularies and to 0.3 for the city name vocabulary.
Fig. 1 shows the number of errors for the digits vocabulary as a function of α. Using values of α greater than 0.3, the recognition rate improves with respect to the baseline system, based on the HMM scores only $(\alpha = 1)$.
The value of α is estimated using an independent evaluation set. To assess the stability of the α value with respect to different evaluation sets, two subsets of the city name *test set*, including 7183 and 7216 utterances respectively, have been considered. For these subsets the number of errors as a function of α has been evaluated and is plotted in Fig. 2. Since the behavior of the two curves is similar with respect to α, we can be confident that the best value chosen for a subset will be also good for an independent test set collected in the same conditions.

These results show that the rescoring strategy is able, with minimal overhead, to reduce the error rate of all the tested systems. It is worth noting, however, that in these experiments the word endpoints where automatically detected, but manually controlled.

To test the robustness of the temporal models with respect to automatically detected endpoints an experiment has been performed using the 475 city name database. Head and tail HMMs were trained to account for possible noise or extra linguistic segments at the beginning and ending of words. Each word was modeled, thus, by its sequence of HMM sub-word units including optional leading and trailing noise models.

During recognition, Viterbi alignment detects for each word candidate its endpoints *excluding the initial and*

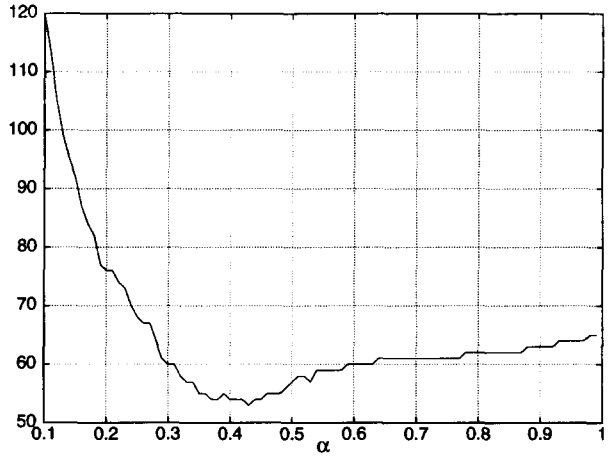

Figure 1: Errors for the digits task as a function of α

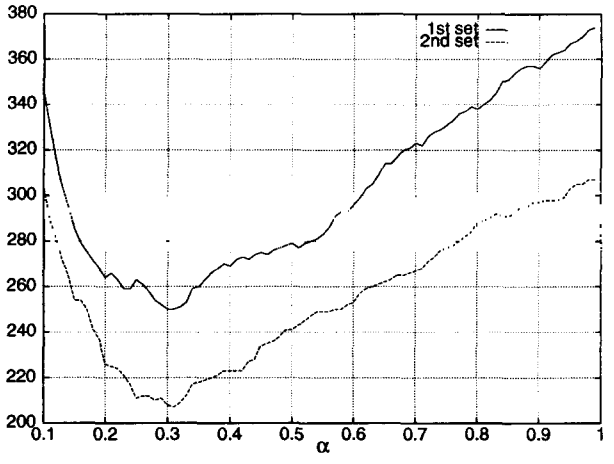

Figure 2: Errors for two subsets of the city name task as a function of α

final noise, that are covered by the head and tail models. Notice that these boundaries may differ for each word candidate, thus, more than one cepstral-time matrix is possibly computed to obtain the temporal scores. This approach, however, does not introduce an appreciable overhead for the computation of the temporal DCTs because our candidate list is limited to a maximum of 10 words, and many word hypotheses have the same endpoints.

The results of this experiment, reported in the fourth column of Table 1, show that the substitution errors of the rescoring strategy are lower than the corresponding errors of the baseline system without automatic endpoint detection. The automatic endpoint detection introduces also 82 deletion and 97 insertion errors.

6. CONCLUSIONS

We have presented a simple, yet effective, post processing approach for isolated word speech recognizers with fixed vocabularies. It has been shown that rescoring the hypotheses produced by a HMM recognizer taking into account the global temporal structure of the pronounced words is effective not only for small vocabularies, but also for a system with a medium size vocabulary of 475 city names included in a vocal service that provides information about the main railway connections.
The main drawback of this approach is that it cannot be applied to flexible vocabularies since it needs several utterances of a word to reliably estimate its temporal model. However, many vocal services provided through the telephone line use fixed vocabularies. For these applications, databases with several samples of each word are collected because it is well known that word models trained with application specific samples are more robust than models obtained by the concatenation of application independent subword units.

We plan to use this post processing approach to rescore the N-best hypotheses produced by a recognizer of connected digits collected through the telephone line.

7. REFERENCES

[1] Y. Ariki, S. Mizuta, M. Nagata, T. Sakai "Spoken-Word Recognition using dynamic features analysed by two-dimensional cepstrum", IEE Proceedings, Vol. 136, Pt. I, No. 2, pp. 133-140, April, 1989.

[2] Furui S., "On the Role of Spectral Transition for Speech Perception", Journal of Acoustic Society of America, Vol. 80, No, 4, pp. 1016–1025, 1986.

[3] Milner B.P., Vaseghi S.V., "Speech Modelling using Cepstral-Time Feature Matrices and Hidden Markov Models", Proceedings of Int. Conference on Acoustic Speech and Signal Processing, Adelaide, Vol. I, pp. 601-604, 1994.

[4] Milner B.P., Vaseghi S.V., "An Analysis of Cepstral-Time matrices for Noise and Channel Robust Speech Recognition", Proceedings of EUROSPEECH'95, Madrid, pp. 519–522, 1995.

[5] Ramesh P., Wilpon J., "Modeling State Durations in Hidden Markov Models for Automatic Speech Recognition", Proceedings of Int. Conference on Acoustic Speech and Signal Processing, Vol.1, pp. 381-384, San Francisco, CA, 1992.